Interference coordination in wireless networks: a flow-level perspective

Richard Combes (*,†), Zwi Altman (*) and Eitan Altman (†)
*Orange Labs
38/40 rue du Général Leclerc, 92794 Issy-les-Moulineaux
Email: firstname.lastname@orange.com
†INRIA Sophia Antipolis
06902 Sophia Antipolis, France
Email: firstname.lastname@inria.fr

***Abstract*—In dense wireless networks, inter-cell interference highly limits the capacity and quality of service perceived by users. Previous work has shown that approaches based on frequency reuse provide important capacity gains. We model a wireless network with Inter-Cell Interference Coordination (ICIC) at the flow level where users arrive and depart dynamically, in order to optimize quality of service indicators perceivable by users such as file transfer time for elastic traffic. We propose an algorithm to tune the parameters of ICIC schemes automatically based on measurements. The convergence of the algorithm to a local optimum is proven, and a heuristic to improve its convergence speed is given. Numerical experiments show that the distance between local optima and the global optimum is very small, and that the algorithm is fast enough to track changes in traffic on the time scale of hours. The proposed algorithm can be implemented in a distributed way with very small signaling load.** [1]

***Index Terms*—Wireless Networks;Queuing Theory;Traffic Engineering;Self-Organizing Networks;Stochastic Approximation;Stability;OFDMA;Load Balancing;Self configuration;Self Optimization**

I. Introduction

As wireless networks become increasingly dense to accommodate the rising traffic demand, inter-cell interference becomes one of the limiting factor as far as performance is concerned. Interference can be managed at various layers and time-scales. At the physical layer, two promising approaches are multi-user detection ([1]) and multi-antenna techniques such as beam-forming and network Multiple Input Multiple Output (MIMO). At the MAC layer, inter-cell scheduling is believed to provide large capacity gains ([2]). On a slower time scale, which we consider in this article, approaches based on frequency reuse other than reuse 1 can provide significant performance improvements ([3]). Soft reuse and fractional reuse are two such approaches.

In this article we propose self-optimizing algorithms to automatically tune the parameters of frequency reuse schemes based on measurements. The self-optimization allows to configure the network and adapt it to daily traffic patterns automatically. The proposed algorithms fall within the scope of Self-organizing networks (SON) ([4]). The SON technology is expected to enable complex and costly network management tasks such as node deployment and configuration, parameter optimization and troubleshooting to be performed automatically.

Previous work on self-organizing ICIC in wireless networks ([5], [6]) does not take into account flow-level dynamics i.e users arrival and departures explicitly and adopts a static approach. Namely, it is assumed that the number of active users and their position remain constant for a sufficient amount of time, so that the transmit powers of the Base Stations (BSs) can be adjusted to maximize a utility function of the users throughput. The flow level performance is then assessed through simulation.

For N active users in the network, define D_n as the throughput of the n-th user. A popular approach for ICIC is to adjust transmit powers to maximize the α-fair utility ([7]):

$$\sum_{1\leq n\leq N} \frac{D_n^{1-\alpha}-1}{1-\alpha}. \tag{1}$$

There is a strong interaction between the chosen rate allocation (the value of α) and the congestion process, i.e the stochastic process describing the number and location of active users as a function of time. In the context of wired networks, [8] investigates the topic and suggests that $\alpha = 2$ is appropriate for elastic traffic. Numerical studies in [6] confirm that $\alpha = 2$ is adequate for ICIC in wireless networks with elastic traffic. The problem with this utility-based approach is that the optimal value of α is not known a priori, making the design complex.

Our contribution : Our contribution is to propose a new approach: we model the system at the flow level taking into account users arrivals and departures, and optimize directly functions of the loads of stations. All flow-level performance indicators such as blocking rate and file transfer time can be expressed as function of the loads, so this approach allows to prove mathematically the convergence of the proposed mechanism to a configuration which is optimal at the flow level. This flow level approach to SON was used in [9] to adjust cell sizes automatically based on network measurements.

We propose an algorithm to tune the parameters of ICIC schemes automatically based on measurements. The conver-

[1]This work has been partially carried out in the framework of the FP7 UniverSelf project under EC Grant agreement 257513

gence of the algorithm to a local optimum is proven, and a heuristic to improve its convergence speed is given. Numerical experiments show that the distance between local optima and the global optimum is very small, and also that the algorithm is fast enough to track changes in traffic on the time scale of hours. The proposed algorithm can be implemented in a distributed way with very small signaling load.

The remainder of the paper is organized as follows: in section II we state the traffic model and the calculation of the network flow-level performance. In section III we expose the three main ICIC schemes that have received attention in the literature, and explain how to calculate user data rates with each of them. In section IV we present a general algorithm to tune the parameters of the ICIC schemes, we prove its convergence and demonstrate that it can be implemented in a distributed way with very little signaling load. In section V we illustrate the performance of the proposed scheme numerically. Section VI concludes the paper.

II. Model for elastic traffic

We write $A \subset R^2$ the network area. There are N_s BSs, and we define A_s the area covered by BS s. Users arrive in the network according to a Poisson process on $A \times R$ of measure $\lambda(dr \times dt) = \lambda(dr) \times dt$. Namely the mean number of users who arrive in the network during time $[t_1, t_2)$ in any (Borel) region $A' \subset A$ is equal to $(t_2 - t_1)\lambda(A')$.

The data rate of a user located at $r \in A$ who is attached to BS s and is the only user served by s is denoted by $R_s(r)$. Elastic traffic is appropriate for modeling File Transfer Protocol (FTP)-like services. Users want to download a file of size σ, with $E[\sigma] < +\infty$ and $E[\sigma^2] < +\infty$. Round-Robin scheduling is used: when BS s has n_s active users, the throughput of a user located at r is $\frac{R_s(r)}{n_s}$. Each station can be modeled by a M/G/1/PS(Processor sharing) queue ([10]). The load of BS s is:

$$\overline{\rho}_s = E[\sigma] \int_{A_s} \frac{\lambda(dr)}{R_s(r)}, \tag{2}$$

and BS s is stable if $\overline{\rho}_s < 1$ i.e the distribution of the number of active users tends to a finite stationary distribution. The average number of active users in s in stationary state is:

$$E[n_s] = \frac{\overline{\rho}_s}{1 - \overline{\rho}_s} \tag{3}$$

Using Little's law ([11]), the mean file transfer time in the network is given by the expected number of active users divided by the arrival rate:

$$E[F] = \frac{1}{\lambda(A)} \sum_{s=1}^{N_s} \frac{\overline{\rho}_s}{1 - \overline{\rho}_s}. \tag{4}$$

With admission control the blocking rate is:

$$B_s = \frac{\overline{\rho}_s^{N_{max}}}{\sum_{i=1}^{N_{max}} \overline{\rho}_s^i} \tag{5}$$

with N_{max} - the maximal allowed number of active users in a BS.

III. Interference coordination schemes

In order to perform ICIC, we must show the link between the powers transmitted by BSs and the data rates $R_s(r)$. The data rate calculation is done for several ICIC schemes that have received attention in the literature.

While this article is written with Orthogonal Frequency-Division Multiple Access (OFDMA) networks in mind, the results hold for any access scheme where the radio resources can be divided in a set of parallel orthogonal channels. The term "resources" stands for: time-frequency blocks for OFDMA, codes for Code Division Multiple Access (CDMA) (when inter-code interference can be ignored), and time slots for Time Division Multiple Access (TDMA). We denote by N_r the number of resources, $h_s(r)$ - the signal attenuation comprising path loss and shadowing between BS s and location r, and ϕ - the mapping between Signal to Interference plus Noise Ratio (SINR) and data rate on a resource, for an Additive White Gaussian Noise (AWGN) channel. In subsequent sections, we will choose ϕ as:

$$\phi(S) = bW \log_2(1 + \frac{S}{a}), \tag{6}$$

with W being the bandwidth for a resource, and $a \geq 1$, $b \leq 1$ two constants. For $a = 1$ and $b = 1$, (6) is simply the Shannon formula. a represents the loss of efficiency due to practical (finite length) coding schemes, and b the proportion of effectively usable bandwidth, since part of the bandwidth is used for signaling in practical systems. A very good fit between (6) and link-level simulations was shown by [12], and suggested that $a = 1.25$ and $b = 0.75$ were the correct values for Long Term Evolution (LTE) systems.

We define $Z_s(r)$ the amplitude of the channel fading between BS s and location r on a resource, with $E[Z_s(r)] = 1$. We assume independence of the fading processes on different resources. We further assume independence of the fading across BSs, i.e $Z_s(r) \perp\!\!\!\perp Z_{s'}(r)$, $s \neq s'$. The coherence time of the channel is much smaller than the time scale on which users arrive and depart, so that the throughput of a user in a given user configuration can be considered equal to the expected data rate, when expectation is taken on the fading.

A. Fractional load

1) Data rate calculation: The first scheme is denoted Fractional Load (FL), as introduced in [13]. Instead of using all available resources, BS s uses each resource with probability $0 < p_s \leq 1$. This enables reducing inter-cell interference since the average interference caused by BS s to its neighbours is proportional to p_s. This scheme is completely decentralized, since each BS chooses the resources used for transmission independently of the decision taken by other BSs. If a BS uses a resource, it transmits at a fixed power P_{max}. We define a random variable $U_s \in \{0, 1\}$, with $P[U_s = 1] = p_s$, and $U_s \perp\!\!\!\perp U_{s'}$, $s \neq s'$, since BSs take their decisions independently. Let $S_s(r)$ denote the SINR between BS s and

location r on a resource, when it is used:

$$S_s(r) = \frac{P_{max} h_s(r) Z_s(r)}{N_0^2 + \sum_{s' \neq s} P_{max} h_{s'}(r) U'_s Z_{s'}(r)}, \quad (7)$$

with N_0^2 the thermal noise power on a resource. The total data rate is proportional to the number of used resources:

$$R_s(r) = N_r E\left[U_s \phi(S_s(r)\right] = N_r p_s E\left[\phi(S_s(r))\right] \quad (8)$$

The expression in (8) can be calculated by BS s since:

- P_{max} is known
- Channel attenuations $h_s(r)$, $\forall s$, the distribution of the channel fading $Z_s(r)$, $\forall s$ and the resource utilization p_s, $\forall s$ can be known through measurements done by the users.

2) Simplified calculation: Although the data rate in (8) can be calculated, it involves a fairly large amount of signaling between users and the BSs, in particular to know the channel statistics $Z_s(r)$, $\forall s$. In practical scenarios, the calculation can be made much simpler as long as we can assume that there are a large number of interfering BSs. In this case we can calculate the data rate (8) replacing the interference by the mean interference. [14] shows that for hexagonal networks in a urban environment, the approximation is very accurate.

We write $I_s(r)$ the mean interference at location r when served by station s:

$$I_s(r) = P_{max} \sum_{s' \neq s} p_{s'} h_{s'}(r) \quad (9)$$

Using Jensen's inequality we obtain a lower bound for the data rate:

$$R_s(r) \geq N_r p_s E\left[\phi(\frac{P_{max} h_s(r) Z_s(r)}{N_0^2 + I_s(r)})\right], \quad (10)$$

which is considerably simpler to calculate than (8) since it does not involve the distribution of the fading of interfering signals $Z_{s'}(r)$, $s' \neq s$.

B. Fractional frequency reuse

1) Data rate calculation: Another scheme is Fractional Frequency Reuse (FFR), which is based on frequency planning and is considered in [5] as a basis for self-organizing ICIC. Resources are divided into N_b sub-bands of equal size. The BSs use all the resources, but they do not transmit the same power on all sub-bands. Namely, if a BS transmits at strong power on a given sub-band, then its neighbors should transmit at smaller power in order to avoid creating too much inter-cell interference. Reuse patterns appear in the network, which enables mitigating the inter-cell interference. Each user receives each resource of each sub-band an equal amount of time, i.e Round-Robin scheduling applies.

Let $P_{s,b}$ denote the power transmitted by BS s on a resource of sub-band b and $S_{s,b}(r)$ - SINR at location r, on a resource of sub-band b when served by BS s:

$$S_{s,b}(r) = \frac{P_{s,b} h_s(r) Z_s(r)}{N_0^2 + \sum_{s' \neq s} P_{s',b} h_{s'}(r) Z_{s'}(r)} \quad (11)$$

The data rate is:

$$R_s(r) = \frac{N_r}{N_b} \sum_{1 \leq b \leq N_b} E\left[\phi(S_{s,b}(r))\right]. \quad (12)$$

2) Simplified calculation: As said in III-A2, the data rate calculation in (12) can be simplified. The mean interference at r when served by BS s on a resource of sub-band b is:

$$I_{s,b}(r) = \sum_{s' \neq s} P_{s',b} h_{s'}(r). \quad (13)$$

The simplified expression for the data rate is:

$$R_s(r) \geq \frac{N_r}{N_b} \sum_{1 \leq b \leq N_b} E\left[\phi(\frac{P_{s,b} h_s(r) Z_s(r)}{N_0^2 + I_{s,b}(r)})\right]. \quad (14)$$

C. Soft frequency reuse

1) Data rate calculation: The last scheme considered is called Soft Frequency Reuse (SFR), and performance studies in [3] show that it enables a significant increase in capacity in dense networks. Region A_s served by BS s is divided into two regions denoted "center" and "edge" $A_{s,c}$ and $A_{s,e}$, based on a path-loss threshold, i.e users far from the BS are called edge users and the other are called center users. [3] further shows that it is optimal to choose the path-loss threshold as the median path-loss in the cell. Resources are divided in 3 sub-bands of equal size. One sub-band is used to serve edge users on which BS s transmits at maximal power P_{max}, and two sub-bands are used for center users on which it transmits at low power $P_{max}\kappa_s$. Typical values for κ_s are around $-10dB$ ([3]). We define $\mathcal{B}_{s,e}$ the sub-band used by BS s to serve edge users, and $\mathcal{B}_{s,c}$ the set of two sub-bands used by BS s to serve center users. Using the same notations as for FFR, we have that $P_{s,b} = P_{max}$ if $b \in \mathcal{B}_{s,e}$ and $P_{s,b} = P_{max}\kappa_s$ if $b \in \mathcal{B}_{s,c}$. The SINR $S_{s,b}(r)$ is calculated as for FFR by (11).

The data rate for edge users $R_{s,e}(r)$ is:

$$R_{s,e}(r) = \frac{N_r}{N_b} \sum_{b \in \mathcal{B}_{s,e}} E\left[\phi(S_{s,b}(r))\right], \quad (15)$$

and the data rate for center users $R_{s,c}(r)$ is:

$$R_{s,c}(r) = \frac{N_r}{N_b} \sum_{b \in \mathcal{B}_{s,c}} E\left[\phi(S_{s,b}(r))\right]. \quad (16)$$

The previous remark III-A2 on the simplified data rate calculation remains valid.

2) Queuing model: Each station can be modeled as 2 parallel M/G/1/PS queues ([3]) with loads:

$$\overline{\rho}_{s,e} = E\left[\sigma\right] \int_{A_{s,e}} \frac{\lambda(dr)}{R_{s,e}(r)}, \quad (17)$$

$$\overline{\rho}_{s,c} = E\left[\sigma\right] \int_{A_{s,c}} \frac{\lambda(dr)}{R_{s,c}(r)}. \quad (18)$$

As previously, the mean number of active users in BS s is:

$$E\left[n_s\right] = \frac{\overline{\rho}_{s,e}}{1 - \overline{\rho}_{s,e}} + \frac{\overline{\rho}_{s,c}}{1 - \overline{\rho}_{s,c}}. \quad (19)$$

IV. Self-organizing interference coordination

In order to compute $\overline{\rho}_s$, we need to know the data rates at every point of the cell $R_s(r)$ and the arrival intensity $\lambda(dr)$. To compute the data rate $R_s(r)$, using the calculations of Section III, we need to know the signal attenuations from all the BSs $h_s(r)$, $1 \leq s \leq N_s$. In practice, this information is not available. In this section, we show that the loads can be estimated by observing users arrivals and file sizes. This allows to optimize a function of the loads in an "online" fashion: the characteristics of the arrival process are not known and the system is optimized based on successive observations. The observations are by nature noisy, and we will show that an optimum can still be found using stochastic approximation theorems.

A. Load estimation

We denote by $\{T_n, r_n, \sigma_n\}_{n\in Z}$ the marked point process of users arrivals, locations and file sizes. Namely T_n is the instant at which the n-th user arrives, r_n his location of arrival, and σ_n its file size. Time is divided in time slots of size T, and the k-th time slot is $[kT, (k+1)T)$. We assume that T is larger than a typical flow duration. We define $\rho_s[k]$ the load estimate for BS s during the k-th time slot:

$$\rho_s[k] = \frac{1}{T}\sum_{n\in Z}\frac{\sigma_n}{R_s(r_n)}\mathbf{1}_{[kT,(k+1)T)}(T_n). \tag{20}$$

We further need to calculate the derivatives of the loads for our optimization algorithm. For the calculations not to depend on the considered ICIC scheme, we denote by θ the parameters of interest. Namely, θ stands for $\{p_s\}_{1\leq s\leq N_s}$ for FL, $\{P_{s,b}\}_{1\leq s\leq N_s, 1\leq b\leq N_b}$ for FFR and $\{\kappa_s\}_{1\leq s\leq N_s}$ for SFR. We define $\nabla_\theta\rho_s[k]$ the gradient with respect to θ of the load estimate for BS s during the k-th time slot:

$$\nabla_\theta\rho_s[k] = -\frac{1}{T}\sum_{n\in Z}\sigma_n\frac{\nabla_\theta R_s(r_n)}{R_s(r_n)^2}\mathbf{1}_{[kT,(k+1)T)}(T_n). \tag{21}$$

The formulas for calculating $\nabla_\theta R_s(r)$ are given in Appendix A. Loads and their derivatives are estimated without bias, as stated by theorem 1. Furthermore, the standard deviation of those estimates is proportional to $\frac{1}{\sqrt{T}}$.

Theorem 1. *(i)* $E[\rho_s[k]] = \overline{\rho}_s(\theta[k])$,
(ii) $E[\nabla_\theta\rho_s[k]] = \nabla_\theta\overline{\rho}_s(\theta[k])$,
(iii) $\mathbf{var}\rho_s[k]$ *and* $\mathbf{var}\nabla_\theta\rho_s[k]$ *are both finite and proportional to* $\frac{1}{T}$.

Proof: see appendix B ∎

It is noted from the expressions of the data rates that $\nabla_\theta\rho_s[k]$ can be computed by BS s provided that it knows θ, the derivative of ϕ, the value of the path-loss at the locations of arrivals of users that arrived during the k-th time slot, and the fading distributions. Once again, from the remark in III-A2, when the number of interfering BSs is large, the interference can be replaced by the mean interference. Then only the distribution of the fading between the serving BS and the users is needed, and the computation is simpler.

B. Load optimization

1) Optimization objective: We write $\overline{\rho} = (\overline{\rho}_1, \cdots, \overline{\rho}_{N_s})$ the load vector. The objective is to minimize a given function of the loads $U(\overline{\rho})$. A case of particular interest is to minimize the average file transfer time, and according to (4), this can be done by choosing

$$U(\overline{\rho}) = \sum_{s=1}^{N_s}\frac{\overline{\rho}_s}{1-\overline{\rho}_s}, \tag{22}$$

where we have ignored the total arrival rate $\lambda(A)$ since it does not depend on θ.

For our proof we will assume that U is differentiable with bounded derivatives. For the file transfer time, we will either assume that the loads are bounded away from 1, or we will replace $f(\overline{\rho_s}) = \frac{\overline{\rho_s}}{1-\rho_s}$ by a smooth function g such that $f = g$ on $[0, 1-d)$ and g is linear on $(1-d, +\infty)$, with d an arbitrarily small constant. Another case of interest is when we try to minimize the load of the most loaded station, i.e $U \equiv \max$. In this case we can use the smooth approximation:

$$U(\overline{\rho}) = \frac{1}{\tau}\log(\sum_{s=1}^{N_s}\exp(\tau\overline{\rho}_s)), \tag{23}$$

with τ a smoothing parameter.

2) Constraint sets: For each ICIC scheme, parameter θ is constrained to a compact convex set $\mathcal{P}$. For FL the constraint set is:

$$\{\mathbf{p}|p_{min} \leq p_s \leq 1\}, \tag{24}$$

where $p_{min} > 0$ since $\overline{\rho}_s \underset{p_s\to 0^+}{\to} +\infty$. For FFR the constraint set is:

$$\{\mathbf{P}|P_{min} \leq \sum_{1\leq b\leq N_b} p_{s,b} \leq P_{max}, 1 \leq s \leq N_s\}, \tag{25}$$

with P_{max} the maximal allowed total transmit power of a BS and $P_{min} > 0$ since $\overline{\rho}_s \underset{\sum_b P_{s,b}\to 0^+}{\to} +\infty$. For SFR the constraint set is:

$$\{\kappa|\kappa_{min} \leq \kappa_s \leq 1\}, \tag{26}$$

with $\kappa_{min} > 0$ since $\overline{\rho}_{s,c} \underset{\kappa_s\to 0^+}{\to} +\infty$. We define a constraint set for the loads $\mathcal{C} = [0, \rho_{max}]^{N_s}$ for FFR and FL, and $\mathcal{C} = [0, \rho_{max}]^{2N_s}$ for SFR. We define $\pi_{\mathcal{P}}[.]$ and $\pi_{\mathcal{C}}[.]$ the projection on $\mathcal{P}$ and $\mathcal{C}$ respectively. We denote by $\frac{\partial U}{\partial s}$ the partial derivative of U with respect to its s-th parameter.

3) Algorithm: We define the filtered loads $C[k]$ for the k-th time slot:

$$C[k+1] = \pi_{\mathcal{C}}[(1-\delta)C[k] + \delta\rho[k]], \tag{27}$$

with $0 < \delta$ a filtering parameter. We define the update vector for the k-th time slot:

$$Y[k] = \sum_{1\leq s\leq N_s}\nabla_\theta\rho_s[k]\frac{\partial U}{\partial s}(C[k]). \tag{28}$$

θ is updated at each time slot and projected back on the constraint set $\mathcal{P}$:

$$\theta[k+1] = \pi_{\mathcal{P}}[\theta[k] - \epsilon Y[k]]. \quad (29)$$

with $\epsilon > 0$ a small step size. Theorem 2 demonstrates that when $\epsilon, \delta \to 0$ with $\frac{\epsilon}{\delta} \to 0$, the sequence $\{\theta[k]\}_{k\in N}$ converges in distribution to $\mathcal{U}$, the set of local minima of U on the constraint set $\mathcal{P}$. The proof is based on stochastic approximation. The performance gap between local optima and the global optimum will be discussed in Section V, and we will show that it is small.

Theorem 2. *For ρ_{max} large enough, $\{\theta[k]\}_{k\in N}$ converges in distribution to $\mathcal{U}$ when $\epsilon \to 0$, $\delta \to 0$ and $\frac{\epsilon}{\delta} \to 0$. Namely, for all $\beta > 0$:*

$$\limsup_k P\left[d_{\mathcal{U}}(\theta[k]) > \beta\right] \underset{\epsilon,\delta,\frac{\epsilon}{\delta}\to 0}{\to} 0, \quad (30)$$

with $d_{\mathcal{U}}(\theta) = \inf_{u\in\mathcal{U}} \|\theta - u\|$ the distance to set $\mathcal{U}$.

Proof: See appendix C ∎

C. Numerical considerations

While the algorithm (29) can be proven mathematically to converge to a local minimum, its performance can further be improved due to the specificity of the problem considered here. After extensive numerical experiments, we suggest a modification of (29) which makes it much more efficient in practice, especially when traffic is not stationary and the algorithm must be fast enough to "track" the traffic variations. The efficiency will be illustrated numerically in Section V.

When $P_{s,b}$ is small, $\frac{\partial U}{\partial P_{s,b}}$ has a large absolute value and its sign varies quickly which forces us to use a small value of ϵ to avoid instability. However when $P_{s,b}$ is large, $|\frac{\partial U}{\partial P_{s,b}}|$ is close to 0, which causes the algorithm to "get stuck" in regions where $P_{s,b}$ is large for a long time, unless ϵ is large. Hence we suggest to work with powers on a logarithmic scale (in dB) so that the steps at low power will be smaller than the steps taken at high power. Furthermore, instead of taking steps proportional $\nabla_\theta U$, we take steps of constant size. Our modified algorithm can be written:

$$\theta[k+1] = \pi_{\mathcal{P}}[\exp\left(\log(\theta[k]) - \epsilon\, \mathbf{sign}(Y[k])\right)]. \quad (31)$$

where $\mathbf{sign}(x)$ is a vector whose components are the signs of the components of x, and $\mathbf{sign}(0) = 0$. This modification enables very good tracking performance.

D. Distributed implementation and signaling load

Since our algorithm runs in real-time, we must show that it can be implemented in a distributed way, where each BS controls its own parameters, with a small signaling load. Namely, we write θ_s the components of θ which are parameters of BS s. That is: $\theta_s = p_s$ for FFR, $\theta_s = \{P_{s,b}\}_{1\le b\le N_b}$ for FFR and $\theta_s = \kappa_s$ for SFR. Assume that U is additive and can be written:

$$U(\overline{\rho}) = \sum_{s=1}^{N_s} u(\overline{\rho}_s), \quad (32)$$

with u a scalar function.

We define $\mathcal{N}_s$ the set of neigbours of BS s, such that $\frac{\partial \overline{\rho}_s}{\partial \theta_{s'}} = 0$ if $s' \notin \mathcal{N}_s$. This means that, in order to be able to calculate the gradient $\frac{\partial U}{\partial \theta_s}$, BS s only needs to know the derivative of the loads of its neigbours. Hence at each time slot, each neighbor of BS s, BS $s' \in \mathcal{N}_s$ will communicate to BS s the value of $\nabla_s u(\overline{\rho}_{s'})$, and in this way the algorithm can be implemented in a distributed way, with the only assumption that there exists an interface between neighboring BSs. Such an interface exists in LTE (X2 interface).

The amount of signaling (exchanged every T seconds) per base station is proportional to the number of components of θ_s multiplied by the number of neighbours. In practical cases, we will have 3 components for FFR, and 1 for FL and SFR, $T = 60s$ and 6 neighbours (hexagonal network). Assuming that floating numbers are coded on 32 bits, the signaling per BS will be of $\frac{3\times 32\times 6}{60} = 9.6\, bits/s$ which is indeed very small for current networks. It is also noted that since information is exchanged every $60s$ or so, the interface delay will not be problematic, since the typical delay value for the X2 interface is $50ms$.

V. Numerical Experiments

In this section we illustrate the performance of the proposed schemes numerically. We show two main features of our method: the algorithm is fast enough to track the changing traffic on the time scale of hours and that although the algorithm is proven to converge to a *local* optimum, the distance to the global optimum is small.

A. Simulation setting

We perform simulations for a hexagonal network with 12 BSs. In order to avoid introducing border effects, we use a wrap-around, which is equivalent to placing the stations on a torus. The measurement interval length is $T = 60s$. T should be reasonably larger than flow durations (a few seconds) in order to avoid sudden changes of the data rate for active flows. In our simulation, we assume that the arrival rate varies slowly. We want to show that the proposed algorithm is able to adapt to the changing traffic distribution. This is essential in practical settings since traffic intensity and distribution changes on the time scale of hours or so. For our numerical experiments we choose the traffic variation to be sinusoidal:

$$\lambda(dr \times dt) = (\lambda_1 + \lambda_2 \mathbf{1}_{A_1}(r) \sin(\frac{2\pi t}{4}))dr \times dt, \quad (33)$$

with t in hours. The arrival rate is the sum of a uniform traffic which does not vary with time, and a traffic in BS 1 which varies sinusoidally. Namely, there is a "hot spot" in BS 1 which appears and disappears periodically. We will compare the performance of the proposed algorithm to a reference scenario (denoted as "no SON") in which all BSs transmit at full power on all resources. Other simulation parameters are given in Table I.

Simulation parameters	
Network layout	Hexagonal
Antenna type	Omni-directional
Number of base stations	12
Inter-site distance	$500m$
Network Area	$1km \times 1km$
Access technology	OFDMA
Link Model	SISO, AWGN + Rayleigh fading
Number of resource blocks	100
Resource block size	$180kHz$
BS maximal transmit power	$46dBm$
Thermal noise	$-174dBm/Hz$
Path loss model	$128 + 37.6 \log_{10}(d)$ dB, d in km
Shadowing standard deviation	6 dB
Average file size	$10Mbytes$

TABLE I
SIMULATION PARAMETERS

B. Results

On Figure 1, we plot the file transfer time in the network as a function of time for different ICIC schemes. During high traffic periods, the network becomes critically overloaded, which results in a high file transfer time. All ICIC schemes bring some improvement and are indeed able to adapt to the changing traffic. The FFR scheme performs the best and greatly reduces the file transfer time, followed by the SFR and the FL comes in last.

For FFR, the transmitted powers by BS 1 and 2 as a function of time are represented on figures 2 and 3 respectively. BS 1 and BS 2 are neighbors. We first notice that BS 1 transmits most of its power on band 2, while BS 2 transmits most of its power on band 3. Since BSs are neighbors, they should indeed avoid transmitting at strong power on the same band, and this shows that the algorithm creates reuse patterns in the network in a autonomous, self-organizing manner. We also notice that during high traffic periods BS 1 increases its transmitted powers noticeably on bands 1 and 3 in order to serve its users faster and avoid congestion.

For FL, the proportion of used resources by BS 1 and 2 as a function of time is represented on figure 4. During the periods of low traffic, both BSs use their resources fully, while during the high traffic periods, BS 2 uses 80% of its resources in order to create less interference to BS 1 which is overloaded.

For SFR, the edge/center power ratio for BSs 1 and 2 as a function of time is represented on figure 5. At low traffic, both BSs use the same power ratio, and at high traffic, BS 1 transmits at stronger power to serve its users faster, while BS 2 transmits at a weaker power in order to create less interference to BS 1 and help reducing its congestion.

On figure 6, we compare the distance between the global optimum and the local minima found by the proposed algorithm. For each ICIC scheme, we first derive the global optimum of U through a global search heuristic (particle swarm was used here). Then we run the proposed algorithm for 1 hour (i.e 60 iterations if $T = 60s$) a hundred times, each time starting from a random point in the constraint set $\mathcal{P}$. We then plot the cumulative distribution function (c.d.f) of the performance over those 100 trials (the first point of the c.d.f being the global optimum). We can see that for the three ICIC schemes, the performance of the proposed algorithm is very close to the global optimum. This is a very interesting result because it indicates that a simple local search achieves good performance without the need to spend possibly an extensive amount of computing power and signaling for a global search.

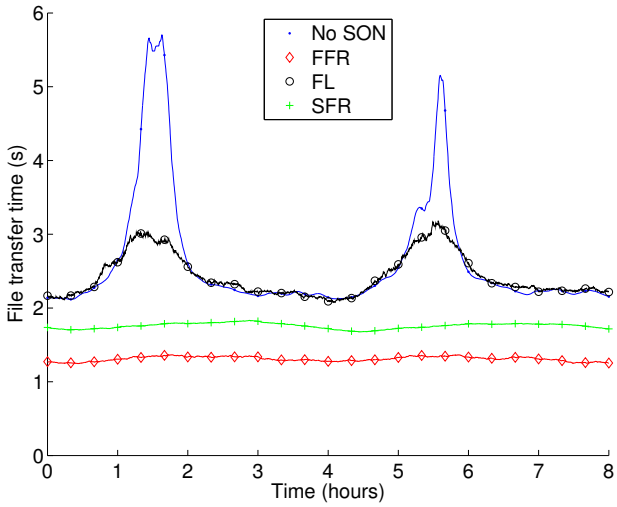

Fig. 1. File transfer time as a function of time, comparison between ICIC schemes

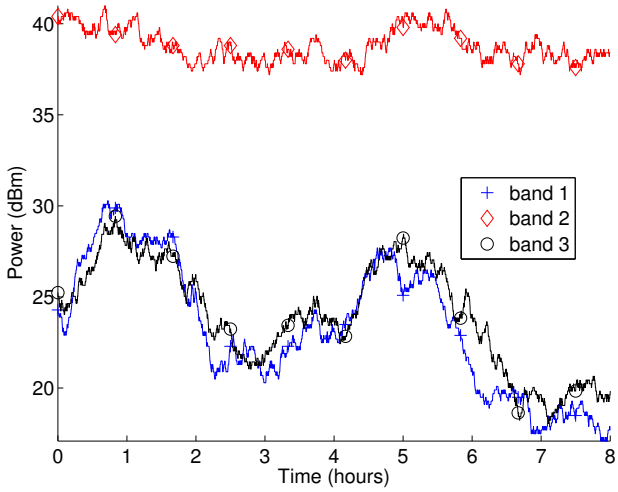

Fig. 2. Powers on different sub-bands transmitted by BS 1 as a function of time for FFR

VI. CONCLUSION

We have considered self-organizing ICIC in wireless networks taking into account flow level dynamics where users arrive and depart dynamically, in order to optimize quality of service indicators perceivable by users such as file transfer time for elastic traffic. We have proposed an algorithm to tune the parameters of ICIC schemes automatically based on measurements. The convergence of the algorithm to a local

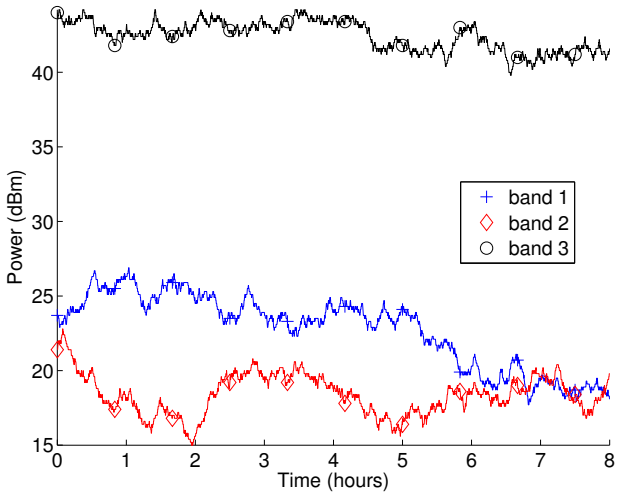

Fig. 3. Powers on different sub-bands transmitted by BS 2 as a function of time for FFR

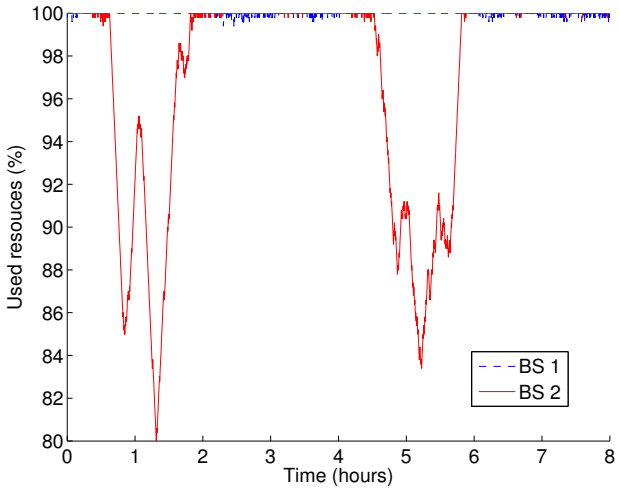

Fig. 4. Resource utilization for BS 1 and 2 as a function of time for FL

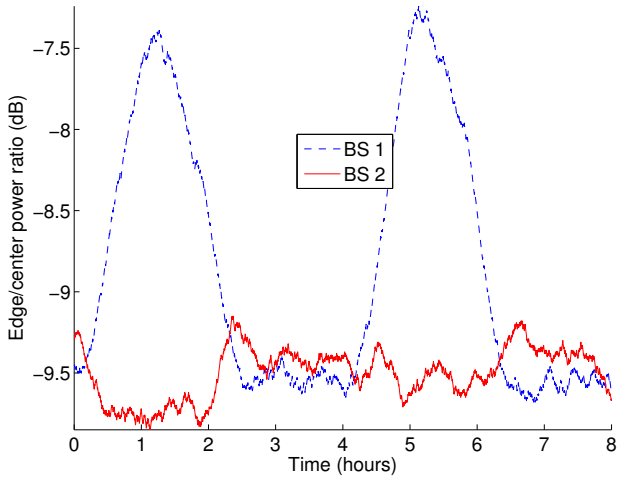

Fig. 5. Edge/center power ratio for BS 1 and 2 as a function of time for SFR

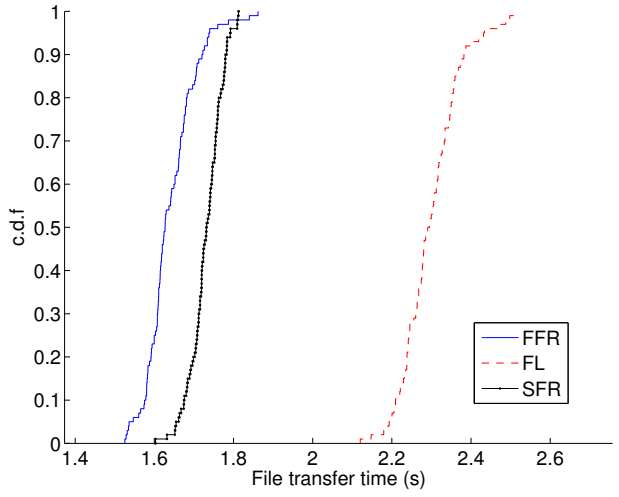

Fig. 6. Comparison of file transfer time achieved through multiple runs for FFR, FL and SFR, comparison between local and global optima

optimum is proven, and a heuristic to improve its convergence speed is given. Numerical experiments show that the distance between local optima and the global optimum is small, and that the algorithm is fast enough to track changes in traffic on the time scale of hours. The proposed algorithm can be implemented in a distributed way with very small signaling load.

APPENDIX A
CALCULATION OF $\nabla_\theta R_s(r)$

We calculate the derivatives of the data rates with respect to the parameters for the three ICIC schemes. We will always use the so-called simplified formulas since they apply in practical scenarios. We denote by ϕ' the derivative of ϕ. The most important message is that BS s can always calculate $\nabla_\theta R_s(r)$ as long as it knows:

- Path loss for both useful signal and interfering signals: $h_{s'}(r)$, $1 \leq s' \leq N_s$,
- The distribution of the fading of the *useful signal* $Z_s(r)$. BS s *does not* need to know the distribution of the fading of interfering signals $Z_{s'}(r)$, $s' \neq s$,
- ϕ' the derivative of the ϕ.

All those quantities are either assumed to be known (e.g the function ϕ') or can be measured and transmitted to BS s by the users it serves.

A. FL

To ease notation we define:

$$\overline{S}_s(r) = \frac{P_{max} h_s(r) Z_s(r)}{N_0 + I_s(r)}. \tag{34}$$

We have that:

$$\nabla_{p_s} R_s(r) = \frac{R_s(r)}{p_s}, \tag{35}$$

and for $s' \neq s$:

$$\nabla_{p_{s'}} R_s(r) = N_r p_s E\left[-P_{max} h_{s'}(r) \frac{\overline{S}_s(r)}{N_0 + I_s(r)} \phi'(\overline{S}_s(r))\right]. \quad (36)$$

B. FFR

To ease notation we define:

$$\overline{S}_{s,b}(r) = \frac{P_{s,b} h_s(r) Z_s(r)}{N_0^2 + I_{s,b}(r)}. \quad (37)$$

We have that:

$$\nabla_{P_{s,b}} R_s(r) = \frac{N_r}{N_b} E\left[\frac{\overline{S}_{s,b}}{P_{s,b}} \phi'(\overline{S}_{s,b})\right], \quad (38)$$

and for $s' \neq s$:

$$\nabla_{P_{s',b}} R_s(r) = \frac{N_r}{N_b} E\left[-h_{s'}(r) \frac{\overline{S}_{s,b}(r)}{N_0^2 + I_{s,b}(r)} \phi'(\overline{S}_{s,b}(r))\right]. \quad (39)$$

C. SFR

The formulas for SFR are deduced from the FFR case.

APPENDIX B
PROOF OF THEOREM 1

Theorem 3. *Consider $\{T_k, r_k\}_{k\in Z}$ a Poisson process on $R^2 \times R$ with measure $\lambda(dr) \times dt$, and marks $\{\sigma_k\}_{k\in Z}$ which are independent and identically distributed (i.i.d) and independent of $\{r_k, T_k\}_k$. Consider $f : R^2 \to R$ positive and measurable, $t_2 > t_1$, and define:*

$$F(t_1, t_2) = \sum_{k\in Z} \sigma_k f(r_k) \mathbf{1}_{[t_1,t_2)}(T_k). \quad (40)$$

Then:

$$E[F(t_1, t_2)] = E[\sigma](t_2 - t_1) \int_{R^2} f(r)\lambda(dr), \quad (41)$$

and:

$$\mathbf{var} F(t_1, t_2) = E\left[\sigma^2\right](t_2 - t_1) \int_{R^2} f(r)^2 \lambda(dr). \quad (42)$$

Proof: We recall the Campbell formula ([15]):

$$E\left[\sum_{k\in Z} g(r_k, T_k)\right] = \int_{R^2\times R} g(r, t)\lambda(dr)dt, \quad (43)$$

with $g : R^2 \times R \to R$ an arbitrary positive measurable function. The first statement is proven by applying the Campbell formula with $g(r, t) = f(r)\mathbf{1}_{[t_1,t_2)}(t)$.

Applying the Campbell formula at the second order we have that:

$$E\left[F(t_1, t_2)^2\right] = E\left[\sigma^2\right](t_2 - t_1) \int_{R^2} f(r)^2 \lambda(dr) + \left((t_2 - t_1) \int_{R^2} f(r)\lambda(dr)\right)^2, \quad (44)$$

so that:

$$\mathbf{var} F(t_1, t_2)^2 = E\left[\sigma^2\right](t_2 - t_1) \int_{R^2} f(r)^2 \lambda(dr), \quad (45)$$

proving the second result. ∎

APPENDIX C
PROOF OF THEOREM 2

We use a two-time scale stochastic approximation argument for proving convergence. The following conditions are true:

- $\{Y[k]\}_{k\in N}$, $\{\rho[k]\}_{k\in N}$ are both uniformly integrable since they are bounded in mean square
- $(\theta, C) \mapsto \sum_{1\leq s\leq N_s} \nabla_\theta \rho_s(\theta) \frac{\partial U}{\partial s}(C)$ and $\theta \mapsto \rho(\theta)$ are continuous
- For a fixed value of θ, the Ordinary Differential Equation (ODE)

$$\dot{C} = \rho(\theta) - C, \quad (46)$$

has a unique globally asymptotically stable point which is $\rho(\theta)$. Once again $\theta \mapsto \rho(\theta)$ is continuous.

The mean ODE for the "slow time scale"is:

$$\dot{\theta} = -\nabla_\theta U(\rho(\theta)), \quad (47)$$

and $\mathcal{U}$ the set of local minima of U is a Lyapunov stable attractor for this ODE. Then applying [16][Theorem 6.1, chap 8, page 287] guarantees that $\theta[k]$ converges to $\mathcal{U}$ in distribution. Namely:

$$\limsup_k P\left[d_{\mathcal{U}}(\theta[k]) > \beta\right] \underset{\epsilon,\delta,\frac{\epsilon}{\delta}\to 0}{\to} 0, \quad (48)$$

for all $\beta > 0$ which concludes the proof.

REFERENCES

[1] S. Verdu, *Multiuser Detection*, 1st ed. New York, NY, USA: Cambridge University Press, 1998.

[2] T. Bonald, S. Borst, and A. Proutiere, "Inter-cell scheduling in wireless data networks," in *Proc. European Wireless*, 2005, pp. 566–572.

[3] T. Bonald and N. Hegde, "Capacity gains of some frequency reuse schemes in OFDMA networks," in *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*, 30 2009-dec. 4 2009, pp. 1 –6.

[4] 3GPP, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions," 3rd Generation Partnership Project (3GPP), TR 36.902, Sep. 2008.

[5] A. Stolyar and H. Viswanathan, "Self-organizing dynamic fractional frequency reuse for best-effort traffic through distributed inter-cell coordination," in *INFOCOM 2009, IEEE*, apr. 2009, pp. 1287 –1295.

[6] R. Combes, Z. Altman, M. Haddad, and E. Altman, "Self-optimizing strategies for interference coordination in OFDMA networks," in *Communications Workshops (ICC), 2011 IEEE International Conference on*, june 2011, pp. 1 –5.

[7] J. Mo and J. Walrand, "Fair end-to-end window based congestion control," *IEEE transactions networking*, vol. 8, pp. 556–566, October 2000.

[8] T. Bonald and L. Massoulié, "Impact of fairness on internet performance," *SIGMETRICS Perform. Eval. Rev.*, vol. 29, no. 1, pp. 82–91, Jun. 2001.

[9] R. Combes, Z. Altman, and E. Altman, "Self-organization in wireless networks: A flow-level perspective," in *INFOCOM, 2012 Proceedings IEEE*, march 2012, pp. 2946 –2950.

[10] T. Bonald and A. Proutière, "Wireless downlink data channels: user performance and cell dimensioning," in *Proceedings of the 9th annual international conference on Mobile computing and networking*, ser. MobiCom '03. New York, NY, USA: ACM, 2003, pp. 339–352.

[11] J. D. C. Little, "A Proof for the Queuing Formula: L= λ W," *Operations Research*, vol. 9, no. 3, pp. 383–387, 1961.

[12] P. Mogensen, W. Na, I. Kovacs, F. Frederiksen, A. Pokhariyal, K. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, "Lte capacity compared to the shannon bound," in *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*, april 2007, pp. 1234 –1238.

[13] A. Pokhariyal, G. Monghal, K. Pedersen, P. Mogensen, I. Kovacs, C. Rosa, and T. Kolding, "Frequency domain packet scheduling under fractional load for the utran lte downlink," in *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*, apr. 2007, pp. 699 – 703.

[14] R. Combes, Z. Altman, and E. Altman, "Scheduling gain for frequency-selective rayleigh-fading channels with application to self-organizing packet scheduling," *Performance Evaluation*, vol. 68, no. 8, pp. 690 – 709, 2011.

[15] D. J. Dalay and D. Vere-Jones, *An introduction to the theory of point processes*. Springer, 2002.

[16] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications 2nd edition*. Springer Stochastic Modeling and Applied Probability, 2003.